# EXHIBIT Y

**Exhibit 17 – Lee**

'156 Patent

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| [**156a**] A device comprising: | Lee's, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit*, discloses a device. Specifically, Lee describes an FPGA-based face detector using neural networks. *See, e.g.*:<br><br>"The study implemented an FPGA-based face detector using Neural Networks and a scalable Floating Point arithmetic Unit (FPU)." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 315 (Abstract).<br><br>"In Section 2, we describe our FPGA implementation of an FPGA-based face detector using neural networks." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 315. |
| [**156b**] at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, | Lee discloses at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value. *See, e.g.*:<br><br>"The FPU provides dynamic range and reduces the bit of the arithmetic unit more than fixed point method does. These features led to reduction in the memory so that it is efficient for neural networks system with large size data bits. The arithmetic unit occupies 39~45% of the total neural networks system area. Therefore bits reduction is needed not only for memory but also for a FPU and system size. Reduction from FPU 32 bits (IEEE 754 single precision) to 16 bits reduced the size of memory and arithmetic units by 50%, having only 1.25% deterioration in the detection rate. In order to determine the least and acceptable bits of the FPU, we examined how representation errors affect a detection rate through the MRRE. The scalable FPU and the error analysis may be useful to determine the details, especially area and speed of FPU for the embedded neural network system." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 315 (Abstract).<br><br>"After face data come in the input node, they are calculated by MAC with weights. Face or non-face is determined by comparing output results and thresholds. For example, if output is |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| | larger than threshold, it means Face. Here, on the FPGA, this determination is decided easily by checking a sign bit after subtraction between output results and threshold." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 316. |
| [**156c**] wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; and | Lee discloses the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input. *See, e.g.*:<br><br>Lee discloses the use of a floating-point format due to its higher dynamic range:<br><br>"Floating point numbers and fixed point numbers are representative number systems to express real numbers. Floating Point Unit (FPU) is slower and bigger than FiXed point Unit (FXU). However FPU provides a good dynamic range which is helpful for neural networks using wide number range with smaller bits. It is also useful to reduce memory having weights [4, 5]." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 315.<br><br>The floating point unit in Lee operates with reduced precision:<br><br>**2.3 Implementation of Reduced Precision FPU**<br><br>Fig. 4 shows the diagram of top module. The module consists of control logic and an FPU arithmetic unit. |

**Exhibit 17 – Lee**

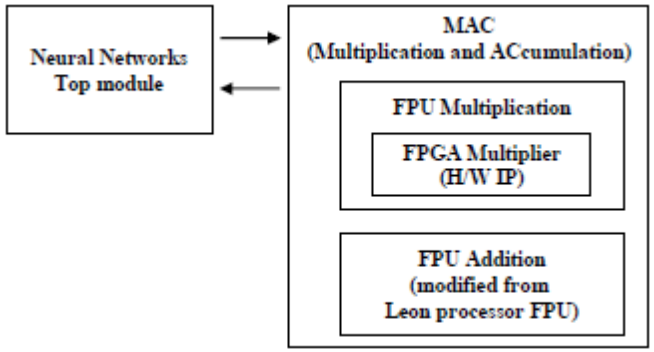
| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| |  <br>Fig. 4 Block diagram of FPU NN<br><br>Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 316-17.<br><br>The reduced precision level of Lee is dependent on the acceptable error range: |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| | **3 Affection of Bit Reduction on FPU**<br><br>**3.1 Representation Errors by FPU Bit Reduction**<br><br>The number of bits in FPU is important to the area and operating speed [8]. Therefore we need to decide the least number of bits within the acceptable error range. We used MRRE [9] as one of the indices of floating point arithmetic accuracy as shown in Table 1. MRRE is the Maximum Relative Representation Error, which is the relative distance between a real number and a represented number. The MRRE can be obtained as follows:<br><br>$$MRRE = \frac{1}{2} \times ulp \times \beta \quad (5)$$<br><br>Where *ulp* is a unit in the last position and $\beta$ is the exponent base.<br><br>Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 317. |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| | Specifically, Lee contemplates a 16-bit floating point unit with six exponent and bit nine mantissa bits:<br>Table 1 MRRE of four Floating Point Units<br>Unit \| $\beta, e, m$ \| Range \| MRRE<br>FPU32 \| 2, 8, 23 \| $2^{2^8-1}=2^{255}$ \| $0.5\square 2^{-23}\square 2=2^{-23}$<br>FPU24 \| 2, 6, 17 \| $2^{2^6-1}=2^{63}$ \| $0.5\square 2^{-17}\square 2=2^{-17}$<br>FPU20 \| 2, 6, 13 \| $2^{2^6-1}=2^{63}$ \| $0.5\square 2^{-13}\square 2=2^{-13}$<br>FPU16 \| 2, 6, 9 \| $2^{2^6-1}=2^{63}$ \| $0.5\square 2^{-9}\square 2=2^{-9}$<br>*See* Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 317.<br>"The FPU representation error increase with MAC calculation of neural networks. The output difference can be expressed as following equations with the notation depicted in Fig. 6. |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|

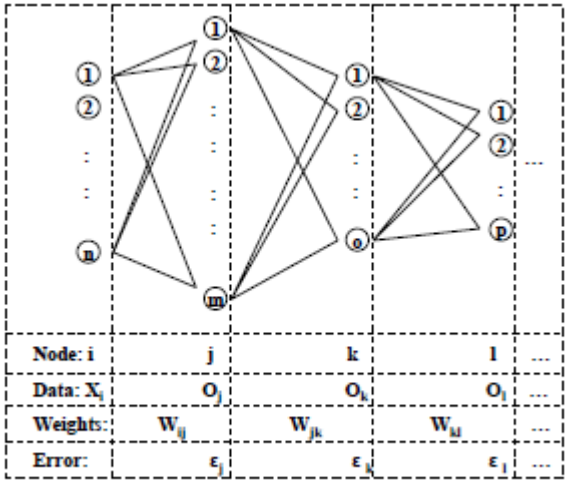



Fig. 6 Neural networks model

The error of the 1st Layer can be described as (6).

$$\varepsilon_j = O^f{}_j - O_j$$
$$= f(\sum_{i=1}^{n} W^f{}_{ij} X^f{}_{ij}) - f(\sum_{i=1}^{n} W_{ij} X_i) + \varepsilon_f . \qquad (6)$$

Where $\varepsilon_j$ represents the hidden layer error, $\varepsilon_k$ represents total error generated between hidden layer and output layer, and $\varepsilon_f$ represents the non-linear function error. $W$ represents the weights and $O$ represents the output of hidden layer.

The term $W^f$ and $X^f$ are new data including the finite precision error. Both are described by

$W^f = W + \varepsilon_W$ and $X^f = X + \varepsilon_X$, respectively.

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| | Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 317.<br><br>After putting the MRRE between FPU32 and other FPU bits into error terms in equations (16) and (18) using the MATLAB and real face data, we can finally find the total accumulated error of neural networks. We considered truncation for a rounding in this paper. Therefore a round-to-nearest scheme will reduce the error more [9].<br>In order to find the detection rate errors, we modeled an EER (Equal Error Rate) graph using Gaussian distribution function and calculated max detection rate error by threshold error variation as shown in Fig. 7.<br><br>Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 318. |

**Exhibit 17 – Lee**

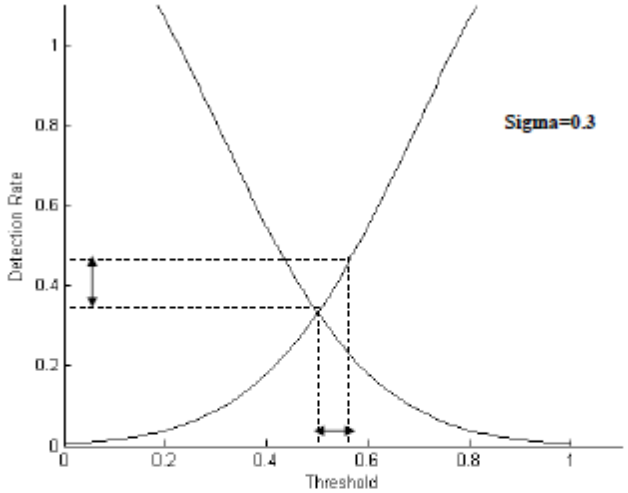
| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| |  <br>Fig. 7 EER graph of detection system<br><br>Bit reduction of 50% from FPU32 to FPU16 is affected by only 1.25 % deterioration in the detection rate as shown in Table 2.<br><br>Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 319. |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|

Table 2 Detection rate errors

| Unit | MRRE | NN OUT Error (Max) | Detection Rate Errors |
|---|---|---|---|
| **FPU32** | $2^{-23}$ | 4.08E-7 | 7.5E-7 |
| **FPU24** | $2^{-17}$ | 2.61E-5 | 4.8E-5 |
| **FPU20** | $2^{-13}$ | 4.18E-4 | 7.7E-4 |
| **FPU16** | $2^{-9}$ | 0.0067 | 0.0125 (1.25%) |

Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 319.

"It is confirmed through MATLAB simulation that this polynomial approximation made about 5 % error in a detection rate." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 316.

"We also showed that bits reduction from FPU 32 bits to 16 bits reduced the size of memory and arithmetic units by 50%, having only 1.25% deterioration in the detection rate." Lee and Ko, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 320.

*See also* Appendix to Responsive Contentions Regarding Non-Infringement and Invalidity ("Responsive Contentions") (detailing error rates associated with different mantissa sizes).

**Exhibit 17 – Lee**

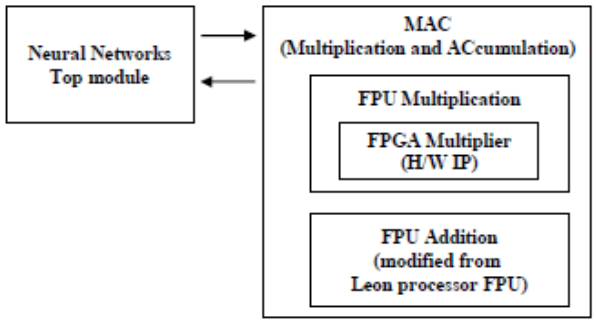
| Claim Limitation (Claim 7) | Exemplary Disclosure |
|---|---|
| [**156d**] at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit; | Lee discloses at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit. *See, e.g.*:<br><br>**2.3 Implementation of Reduced Precision FPU**<br><br>Fig. 4 shows the diagram of top module. The module consists of control logic and an FPU arithmetic unit.<br><br>  <br>Fig. 4 Block diagram of FPU NN<br><br>Lee, *An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit* at 316-17. |
| [**156e**] wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; | To the extent that Singular contends that Lee does not itself identify one of the disclosed computing devices, notwithstanding this disclosure, use of said computing device would have been obvious based on Lee alone or in combination with other disclosed prior art, including without limitation Lee, Belanovic, GRAPE-3, and Cray T3d, for the reasons explained in the Responsive Contentions. |

**Exhibit 17 – Lee**

<table>
<tr><th>Claim Limitation (Claim 7)</th><th>Exemplary Disclosure</th></tr>
<tr><td>[156f] and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.</td><td>“The arithmetic unit occupies 39~45% of the total neural networks system area. Therefore bits reduction is needed not only for memory but also for a FPU and system size. Reduction from FPU 32 bits (IEEE 754 single precision) to 16 bits reduced the size of memory and arithmetic units by 50%, having only 1.25% deterioration in the detection rate.” Lee, An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit at 315 (abstract).<br><br>“Bit reduction of the FPU led to an area reduction and the faster operating clock speed. 50% of bit reduction from FPU 32 to FPU 16 resulted in 50% of addition area (250/486) and memory reduction (1880/3760) as shown in Table 4.” Lee, An FPGA-Based Face Detector Using Neural Network and a Scalable Floating Point Unit at 319.<br><br>Table 4 Analysis of 32, 24, 20, 16-bit of FPUs<br><table><tr><th>Arithmetic Unit</th><th>Memory (Kbits)</th><th>NN Area (Slices)</th><th>FPU Adder Area (Slices)</th></tr><tr><td>FPU 32</td><td>3760</td><td>1077</td><td>486 (45%)</td></tr><tr><td>FPU 24</td><td>2820 (75%)</td><td>878</td><td>403 (45%)</td></tr><tr><td>FPU 20</td><td>2350 (63%)</td><td>750</td><td>300 (40%)</td></tr><tr><td>FPU 16</td><td>1880 (50%)</td><td>650</td><td>250 (39%)</td></tr></table><br>To the extent that Singular contends that Lee does not identify a device with at least 100 multiplication execution units, notwithstanding this disclosure, such a device would have been obvious given intervening FPGA developments for the reasons explained in the Responsive Contentions.</td></tr>
</table>

**Exhibit 17 – Lee**

'273 Patent

| Claim Limitation (Claim 53) | Exemplary Disclosure |
|---|---|
| [**273a**] A device: | Lee discloses a device. *See* [**156a**]. |
| [**273b**] comprising at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, | Lee discloses at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value. *See* [**156b**]. |
| [**273c**] wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X % of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; | Lee discloses the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input. *See* [**156c**]. |
| [**273d**] wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide. | Lee discloses the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide. *See* [**156f**] |

**Exhibit 17 – Lee**

'961 Patent

| Claim Limitation (Claim 4) | Exemplary Disclosure |
|---|---|
| [**961a**] A device comprising: | Lee discloses a device. *See* [**156a**]. |
| [**961b**] at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, | Lee discloses at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a first input signal representing a first numerical value to produce a first output signal representing a second numerical value. *See* [**156b**]. |
| [**961c**] wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=10% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.2% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; and | Lee discloses the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=10% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.2% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input. *See* [**156c**]. |
| [**961d**] at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit. | Lee discloses at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit. *See* [**156d**].<br>. |

**Exhibit 17 – Lee**

| Claim Limitation (Claim 4) | Exemplary Disclosure |
|---|---|
| | |

| Claim Limitation (Claim 13) | Exemplary Disclosure |
|---|---|
| [**961e**] A device comprising: | Lee discloses a device. *See* [**961a**]. |
| [**961f**] a plurality of components comprising: | Lee discloses a plurality of components. *See* [**961b**] + **[961d**]. |
| [**961g**] at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, | Lee discloses at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value. *See* [**961b**]. |
| [**961h**] wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=10% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.2% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input. | Lee discloses the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide. *See* [**961c**]. |